CentraleSupélec

AI Challenge

Presented by :

Tudo Bein

Members of the group :

- João Regazzi
- Guilherme Mertens
- Lucas Tramonte
- Arthur Vogels
- Rebecca Bayssari

Presented to :

- Geraud Faye
- Jean-Philippe Poli

Context

1- Dataset Analysis

2- Overview

3- Pipeline

4- Models

5- Evaluation and results

6- Improvements

Dataset Analysis

DAM training Dataset

- Limited to one image per class
- Uniform Background: Mostly white backgrounds

Test image Dataset

- Diverse backgrounds (different places, stores)
- Real-world variability and complexity

One shot Learning Challenge

- Definition: Learning from just one example per class
- Approach: Train a model to learn semantic representations of the images and compare them with similarity measures

Problem Overview

Train Data ≠ Test Data (Background & Image Capture Differences) →

Preprocessing steps applied

- **Background Removal** – to standardize image conditions
- **Color Retrieval** (For Bags / Shoes / Clothes)

Data Augmentation

1 Image per Class →

Leads to the use of:

- **CNN-based Models with contrastive learning** (e.g., ResNet variants) for feature extraction
- **CLIP Models** for semantic representation of the data

Pipeline

1-Embeddings

Extract feature embeddings from DAM images and test images

Images are transformed into a shared semantic space for comparison space

2-Cosine Similarity

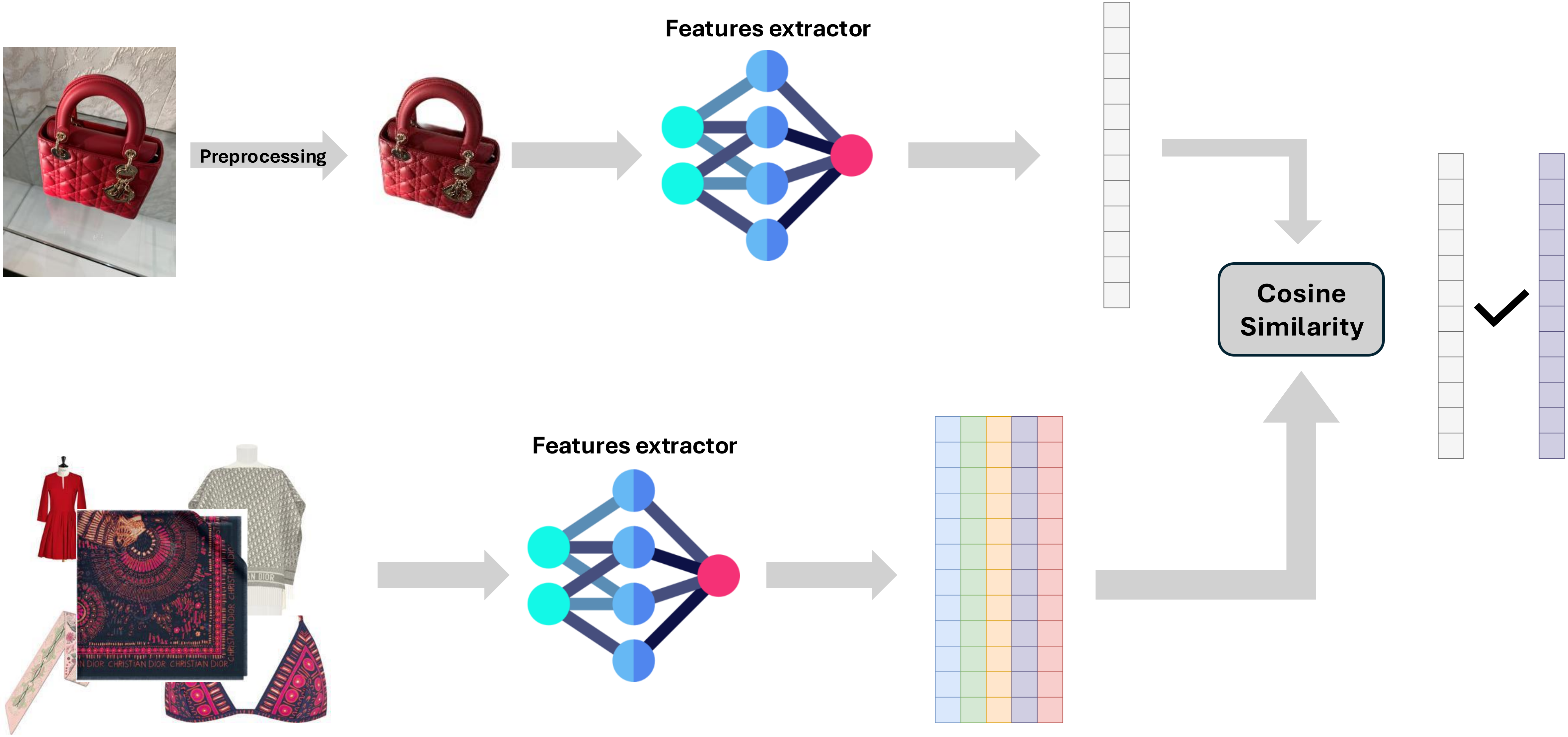
Compute cosine similarity between each test image and DAM reference images.

Measures how visually close the test image is to the reference images from the DAM dataset.

3-Top-5 / Top-10 Matching

Higher-ranked matches have stronger feature similarities

Pipeline



Features extractor
Preprocessing
Cosine
Similarity
Features extractor

Exploring CNN Models for Image Retrieval

1. <u>Image Embedder using pretrained Resnet Backbones</u>

- Test the pipeline with a simple, pretrained model
- Use of Resnet50 pretrained on ImageNet with no finetuning

Those embeddings are not optimized for classification of unseen classes, leading to not so great results

2. <u>Finetuning ResNet with contrastive learning</u>

<u>Contrastive learning:</u> Different images of a same object should have similar embeddings, while images of different objects should have disimilar embeddings.

<u>We compare original images with augmented ones:</u>

The Final approach

CLIP

Source

CLIP was created by OpenAI as an innovative approach to bridge the gap between computer vision and natural language processing

Objective

CLIP's objective in our problem is to learn visual representations from product images to enable accurate image retrieval. By leveraging zero-shot learning, it matches test images to the closest reference in the DAM dataset without requiring task-specific fine-tuning.

Dataset

The model is trained on a massive dataset of 400 million image-text pairs using a self-supervised learning approach

Architecture

CLIP uses a ViT-based transformer to process images by dividing them into fixed-size patches. Its architecture includes multiple transformer layers, attention heads, and feed-forward networks, allowing it to learn rich visual representations for each image

The Final approach

In particular, we use a fine-tuned version of CLIP called:

Fashion Clip*

WHY

- FashionCLIP is designed for real-world applications and is openly available
- It is fine-tuned for the fashion industry, making it highly relevant to our challenge

Architecture

- Uses ViT-B/32 (Vision Transformer) for image encoding.

Training

- Trained on 700k+ fashion image-text pairs (luxury retailers, online stores).
- Fine-tuned to recognize nuanced fashion concepts with zero-shot learning.

* **https://github.com/patrickjohncyh/fashion-clip**

Evaluation Method

Manual Matching

Each test image was manually matched to its corresponding DAM reference image(s)

Some products (e.g., Black Dior Bag) had multiple valid reference images in DAM

FashionCLIP Retrieval

Using the FashionCLIP model, we retrieve the Top 10 most similar images for each test image.

The model ranks results based on cosine similarity.

Evaluation Criteria

A retrieval is correct if the ground truth reference is in the Top 10 results.
Cases with multiple correct references are considered valid.

Average Cosine Similarity Score for correct vs. incorrect matches

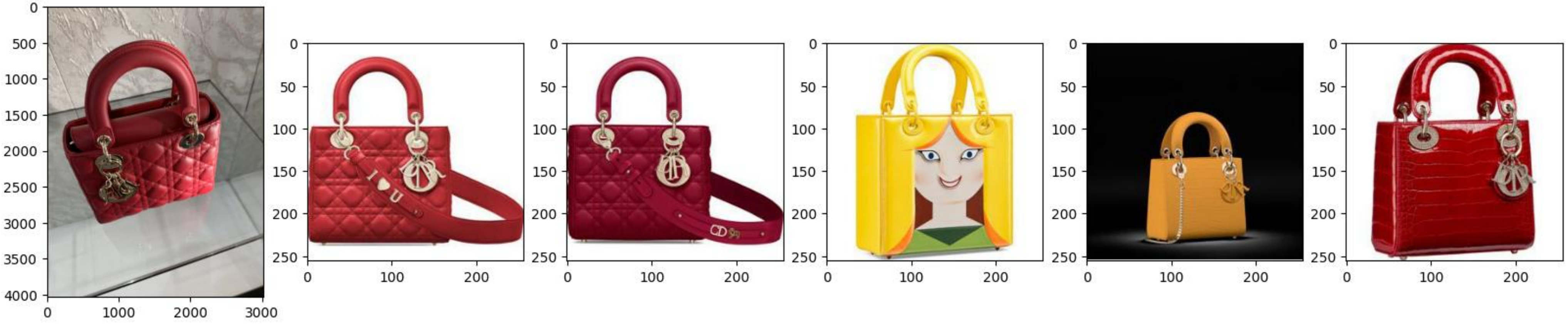
Results

Model	Top 1 accu.	Top 5 accu.	Top 10 accu.
ResNet50	21%	45%	57%
Clip	21%	47%	58%
Fashion-Clip	**47%**	**83%**	**91%**

Comparison of the different models

Qualitative results of the FashionClip model

Results



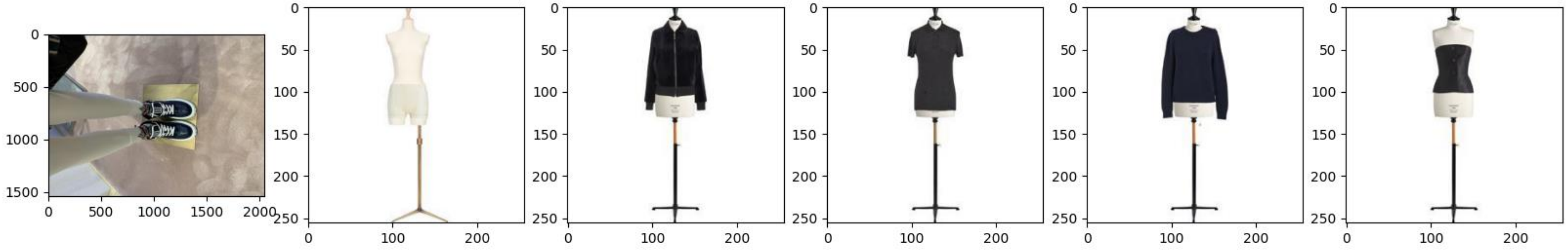
Qualitative results of the ResNet model

Qualitative results of the Clip model

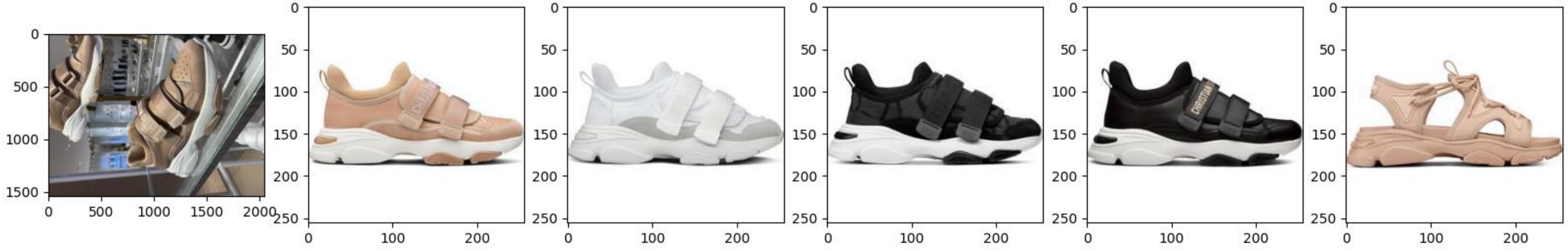
Results

Fashion Clip works well but is not perfect!

Class hallucinations:



Color hallucinations:



Improvements

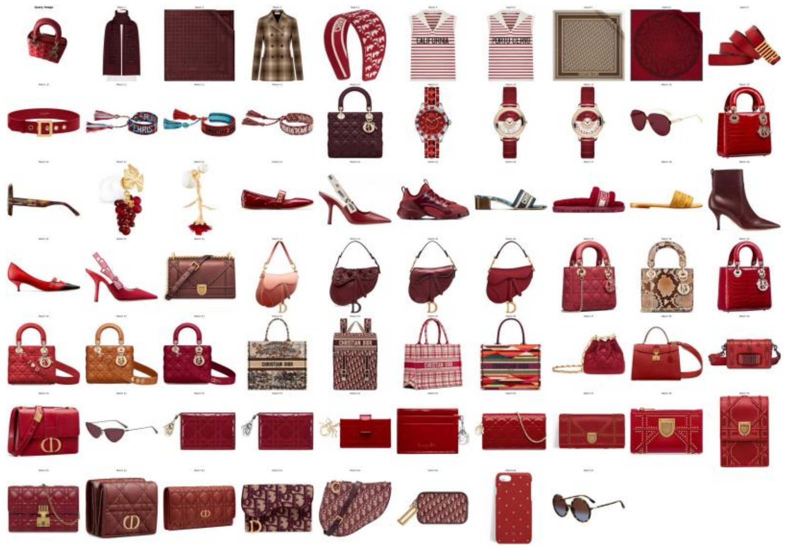
Fashion Clip sometimes misunderstand colors, leading to the idea of adding a color analysis to our pipeline:

If $Distance(A_{colour}, B_{colour}) < threshold$:

$$Similarity(A, B) = \cos(\mathrm{A_{emb}}, \mathrm{B_{emb}}) + \alpha\,(1 - Distance(A_{colour}, B_{colour}) / \, threshold)$$

Else:

$$Similarity(A, B) = \cos(\mathrm{A_{emb}}, \mathrm{B_{emb}})$$



Model	Top 1 accu.	Top 5 accu.	Top 10 accu.
ResNet50	21%	45%	57%
Clip	21%	47%	58%
Fashion-Clip	47%	83%	91%
Fashion-Clip + Color	**51%**	**84%**	**94%**

Conclusion

The best model was :

Fashion Clip with enhanced color analysis

Achievements

- Successfully implemented CLIP for effective image retrieval, leveraging zero-shot learning for unseen classes.
- Achieved over 94% accuracy in top 10 accuracy for the best model.

Future Work

- Experiment with ensemble methods to combine multiple models for improved retrieval accuracy.
- Explore additional augmentation strategies to simulate real-world conditions during training.

References

https://huggingface.co/docs/transformers/en/model_doc/owlv2

https://github.com/NielsRogge/Transformers-Tutorials/blob/master/OWLv2/Zero_and_one_shot_object_detection_with_OWLv2.ipynb

https://pypi.org/project/rembg/

https://huggingface.co/patrickjohncyh/fashion-clip

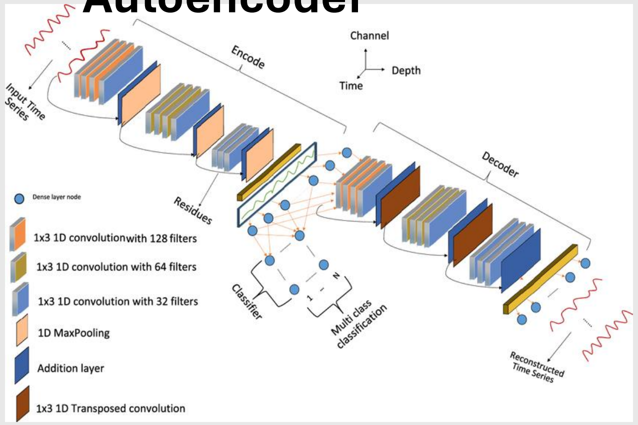
https://youtu.be/oEKg_jiV1Ng?si=_8I9pBSq6BseF86l

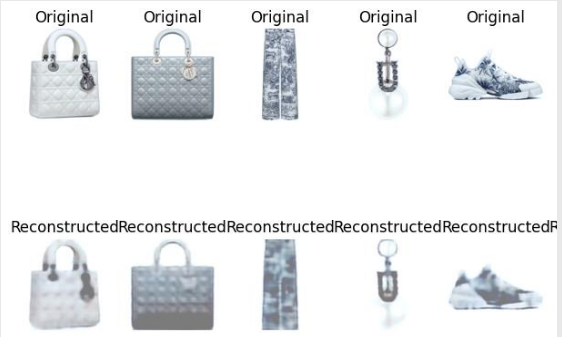
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

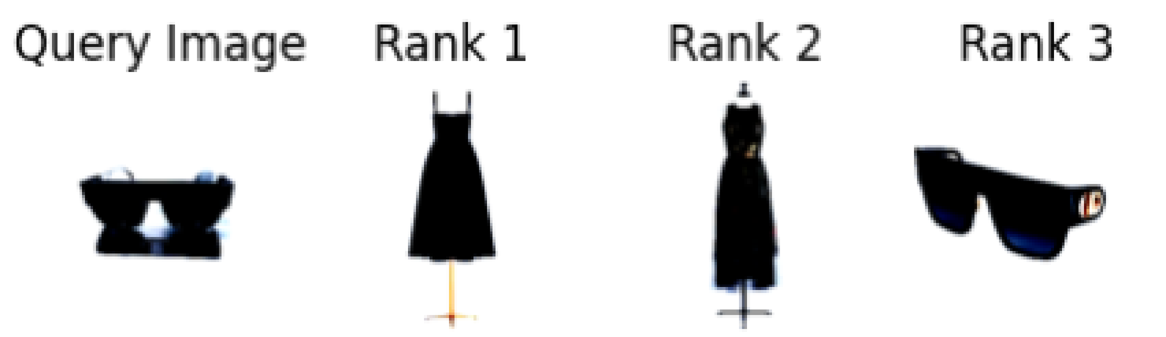
Melekhov, I., Kannala, J., & Rahtu, E. (2016, December). Siamese network features for image matching. In 2016 23rd international conference on pattern recognition (ICPR) (pp. 378-383). IEEE.

Exploring other Models for Image retrieval (ANNEXE)

Autoencoder







OwlViT (Owlv2)

It's a vision transformer model designed for open-world object detection and image retrieval, specifically the "google/owlvit-base-patch32" variant. The model extracts image features (embeddings) using a pretrained transformer architecture, and cosine similarity is then used to match test images to DAM reference images based on visual similarity.

Data Augmentation